

MVP: Winning Solution to SMP Challenge 2025 Video Track

Liliang Ye
Huazhong University of Science and
Technology
Wuhan, China
yll@hust.edu.cn

Yunyao Zhang
Huazhong University of Science and
Technology
Wuhan, China
ikoyun@hust.edu.cn

Yafeng Wu
Huazhong University of Science and
Technology
Wuhan, China
wyf2024@hust.edu.cn

Yi-Ping Phoebe Chen
La Trobe University
Melbourne, Australia
phoebe.chen@latrobe.edu.au

Junqing Yu
Huazhong University of Science and
Technology
Wuhan, China
yjqing@hust.edu.cn

Wei Yang
Huazhong University of Science and
Technology
Wuhan, China
weiyangcs@hust.edu.cn

Zikai Song*
Huazhong University of Science and
Technology
Wuhan, China
skyesong@hust.edu.cn

Abstract

Social media platforms serve as central hubs for content dissemination, opinion expression, and public engagement across diverse modalities. Accurately predicting the popularity of social media videos enables valuable applications in content recommendation, trend detection, and audience engagement. In this paper, we present Multimodal Video Predictor (MVP), our winning solution to the Video Track of the SMP Challenge 2025. MVP constructs expressive post representations by integrating deep video features extracted from pretrained models with user metadata and contextual information. The framework applies systematic preprocessing techniques, including log-transformations and outlier removal, to improve model robustness. A gradient-boosted regression model is trained to capture complex patterns across modalities. Our approach ranked first in the official evaluation of the Video Track, demonstrating its effectiveness and reliability for multimodal video popularity prediction on social platforms. The source code is available at <https://anonymous.4open.science/r/SMPDVideo>.

CCS Concepts

• **Information systems** → **Multimedia information systems**.

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MM '25, Dublin, Ireland

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/2018/06
<https://doi.org/XXXXXXX.XXXXXXX>

Keywords

Social Media Popularity Prediction, Multimodal Machine Learning, Feature Construction

ACM Reference Format:

Liliang Ye, Yunyao Zhang, Yafeng Wu, Yi-Ping Phoebe Chen, Junqing Yu, Wei Yang, and Zikai Song. 2025. MVP: Winning Solution to SMP Challenge 2025 Video Track. In *Proceedings of ACM International Conference on Multimedia 2025 (MM '25)*. ACM, New York, NY, USA, 8 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 INTRODUCTION

In recent years, social media platforms have rapidly evolved into dynamic ecosystems for information dissemination and user interaction, with video content emerging as a dominant medium [7]. The widespread adoption of short videos, live streams, and other multimedia formats has significantly transformed how users express themselves and engage with content [30]. This transformation has led to an unprecedented volume of user-generated videos, posing new challenges and opportunities for understanding content dynamics. Among these, accurately predicting the popularity of social media videos has become a critical task, underpinning key applications such as personalized content recommendation, trend detection, and data-driven decision-making for content creators and platforms alike. However, the popularity of social media videos is influenced by a multitude of factors, including visual features, textual descriptions, user attributes, and contextual information [18, 34, 36]. Effectively integrating multimodal information, constructing expressive feature representations, and leveraging advanced machine learning techniques remain significant challenges in this field.

Recent advances in multimodal learning have integrated visual-language models such as CLIP [21] and VideoMAE [28] with structured context features for video popularity prediction. Contemporary approaches typically combine convolutional neural networks for visual feature extraction with transformer-based architectures

for textual understanding. However, existing methods face critical limitations: most frameworks treat modalities independently during feature extraction, missing crucial cross-modal dependencies; current visual encoders struggle to capture aesthetic and emotional nuances that drive user engagement; temporal modeling approaches often overlook the rapid evolution of trending topics and viral propagation patterns unique to social platforms.

To further advance the understanding of content dynamics in multimodal environments, the Social Media Prediction (SMP) Challenge 2025 introduces a dedicated Video Track centered on short-form user-generated video analysis. The task is defined as a regression problem: given a video post with visual frames, captions, user information, and posting time, the goal is to predict its popularity score. This score reflects aggregated engagement signals and is evaluated using Mean Absolute Percentage Error (MAPE), which accounts for scale variation in real-world data. The dataset, SMPD-Video, includes 6,000 posts from 4,500 users, along with metadata such as categories, tags, and basic video and user attributes. The challenge emphasizes the complex interplay between content, user behavior, and temporal context. By establishing a standardized benchmark, the SMP Challenge 2025 Video Track fosters progress in multimodal learning for social media applications and offers a rigorous testbed for evaluating video-based popularity prediction systems.

In this study, we propose **Multimodal Video Predictor (MVP)**, a robust framework for social media video popularity prediction. Our method leverages a pretrained XCLIP model to extract deep visual representations from sampled video frames, capturing high-level semantic cues. These features are combined with structured metadata, including user statistics, posting time, and content attributes, which are carefully engineered and log-transformed to improve stability. To mitigate the impact of noise, we apply outlier removal and normalization techniques during preprocessing. A CatBoost [8] regressor is then trained on the fused feature set to model complex cross-modal interactions. Our solution achieves top performance on the SMPD-Video dataset, ranking **first place** in the Video Track of the SMP Challenge 2025. This result demonstrates the effectiveness of integrating visual content and structured context for accurate popularity prediction.

2 RELATED WORK

Popularity prediction for visual content originated in image-based analysis [15, 22, 26], where models used features such as color composition and caption text to estimate engagement with tree-based regressors like XGBoost [5, 11, 17, 29, 32, 33]. These early studies demonstrated the value of combining heterogeneous modalities and inspired extensions to video popularity prediction.

With the advent of video content [10, 37], researchers introduced multimodal frameworks [12, 24, 25] to address the inherent noise and uncertainty [23] present in video data. For instance, stochastic embeddings combined with a product-of-experts encoder have been utilized to seamlessly integrate video, textual, and metadata information, thereby improving prediction robustness [19, 38]. Other methods have focused on enhancing representation by retrieving similar videos and incorporating their features during inference, which has shown further improvements in predictive

accuracy [13, 14]. Building upon this retrieval paradigm, recent advances have introduced more sophisticated graph-based approaches that exploit similar-content networks to enhance post representations [6]. These retrieval-augmented modeling techniques leverage external knowledge bases to enrich feature representations with contextual information, while hypergraph-based methods capture complex multi-relational dependencies between users, content, and engagement patterns. The extraction of deep frame-level features from state-of-the-art pretrained video backbones, such as TimeSformer [4], ViViT [2], VideoMAE [28], and X-CLIP [16], has become increasingly prevalent. A representative approach combines such embeddings with BERT-encoded captions and structured metadata, followed by ensemble learning through neural networks and XGBoost to enhance predictive performance [13]. Concurrently, researchers have explored user bias disentanglement techniques to improve model generalization across diverse user populations [9], addressing the challenge of personalized engagement patterns that vary significantly across different user demographics and behavioral clusters. At the same time, vision-language methods transform video frames and captions into descriptive text tokens, using language models for interpretable engagement estimation [20, 27, 31]. Despite these advances, several challenges persist. The highly skewed distribution of video popularity, the presence of noisy or incomplete metadata, and the dynamic nature of user engagement complicate model training and evaluation. The growing importance of structural adaptation and social context modeling in popularity prediction reflects the need for more sophisticated approaches that can handle the complex interplay between content characteristics and social dynamics.

Across these diverse methodologies, several common threads emerge: the integration of multimodal features, the adoption of powerful pretrained representation backbones, and the application of tree-based or hybrid ensemble models. These insights have directly informed the design and development of our MVP framework, which seeks to unify these best practices for robust and interpretable video popularity prediction.

3 METHODOLOGY

3.1 Overview

In this section, we present the design of the Multimodal Video Predictor (MVP) framework, which systematically addresses the challenge of predicting video popularity in social media contexts. As illustrated in Figure 1, the architecture of MVP is tailored to integrate heterogeneous sources of information and model complex interactions across multiple modalities. The pipeline begins with the extraction of high-level semantic representations from video content using a pretrained visual encoder. These visual embeddings are then concatenated with structured features derived from user profiles, temporal signals, and post metadata. To ensure feature compatibility and minimize the influence of noise, we implement a series of preprocessing steps, including log transformation, normalization, and outlier removal. The resulting multimodal feature set is subsequently used to train a gradient-boosted regression model, which is capable of capturing intricate cross-modal dependencies for accurate popularity estimation. The overall framework is designed to be robust and interpretable, facilitating reliable modeling

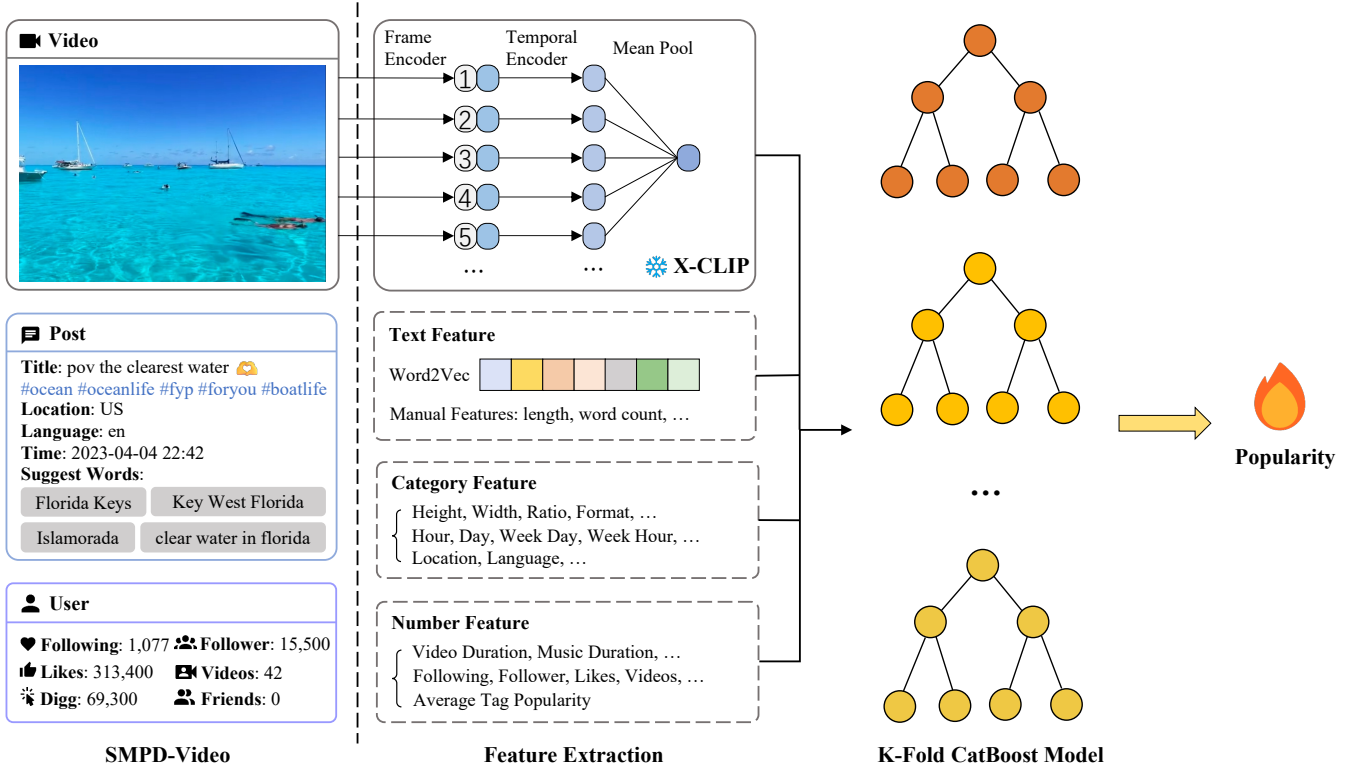


Figure 1: The MVP framework pipeline processes semantic visual features, user profiles, temporal signals, and metadata for multimodal prediction of video popularity in social media using a gradient-boosted regression model.

of the diverse and dynamic nature of social media video data. In the following subsections, we describe each component of the pipeline in detail.

3.2 Multimodal Feature Construction and Fusion

A central aspect of the MVP framework is the systematic extraction and integration of multimodal features that capture the diverse factors influencing social media video popularity. We organize feature engineering into three principal components: visual features, user-related features, and temporal-metadata features, each offering distinct perspectives on user-generated content.

Visual Feature Extraction. We employ a pretrained XCLIP encoder to extract deep visual representations from each video. For every video, we uniformly sample a fixed number of frames to ensure temporal coverage. We process each frame with the XCLIP model, generating high-dimensional embeddings that represent semantic cues such as scene context, objects, and actions. After obtaining these frame-level embeddings, we apply average pooling to aggregate information across frames, resulting in a compact video-level feature. To further balance expressiveness and computational efficiency, we use Principal Component Analysis (PCA) [1] to reduce the dimensionality of the stacked embeddings.

Specifically, let $\mathbf{v}_i \in \mathbb{R}^d$ denote the embedding of the i -th sampled frame, and N be the total number of frames. The video-level feature

\mathbf{v}_{video} is computed by average pooling:

$$\mathbf{v}_{video} = \mathbf{W}^T \left(\frac{1}{N} \sum_{i=1}^N \mathbf{v}_i - \boldsymbol{\mu} \right) \quad (1)$$

where \mathbf{W} is the PCA projection matrix and $\boldsymbol{\mu}$ is the mean vector. This process yields a compact and expressive feature for each video, facilitating robust downstream modeling.

User-Related Features. To characterize the influence of users, we extract detailed user profile statistics. These features include follower count, following count, total published videos, cumulative likes, diggs, hearts, friends, and the average popularity of historical posts for each user. We apply logarithmic transformation to all user-related counts to reduce skewness and enhance numerical stability. These features capture both the social reach and historical activity of the user, which are essential for understanding content dissemination.

Temporal, Metadata, and Semantic Features. To capture periodic engagement patterns, we encode posting time using multiple categorical features, such as hour of day and day of week. Each post's metadata includes video category, tags, language, location, and video-specific attributes like resolution, aspect ratio, duration, and music identifiers. Further feature enrichment involves processing video captions and suggested keywords with standard natural language processing steps, including lowercasing, tokenization, and

embedding extraction using a pretrained Word2Vec model. This process yields additional features, such as caption length, token count, and the number of suggested keywords. Additionally, computing the mean popularity of tags for each post provides information about topic trends and audience preferences. Finally, label encoding or embedding for categorical metadata and normalization of all continuous features ensure consistency across modalities.

Feature Fusion and Preprocessing. We concatenate all extracted features to form a unified multimodal feature vector for each post. Formally, let \mathbf{v}_{visual} , \mathbf{v}_{user} , and \mathbf{v}_{meta} represent the visual, user-related, and temporal-metadata feature vectors, respectively. The final multimodal feature vector \mathbf{v}_{multi} is constructed as:

$$\mathbf{x} = \mathbf{v}_{multi} = [\mathbf{v}_{visual}; \mathbf{v}_{user}; \mathbf{v}_{meta}] \quad (2)$$

where $[\cdot; \cdot; \cdot]$ denotes the concatenation operation. Before feeding the features into the model, we impute missing values with context-appropriate defaults, filter outliers using the interquartile range, and normalize continuous variables. To ensure reliable regression, we preprocess the training labels by removing outliers outside the $[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR]$ range, where Q_1 and Q_3 denote the first and third quartiles, and IQR is the interquartile range. This approach ensures stable feature distributions and minimizes the impact of noisy or extreme values, supporting robust model training. By combining deep semantic visual descriptors, comprehensive user and contextual metadata, temporal signals, and rich textual features, we create a unified and expressive representation for downstream regression modeling. This integrated approach enables the MVP framework to effectively capture the multifaceted drivers of video popularity in social media environments.

3.3 Regression Model

To estimate the popularity score of each video post, we employ CatBoost as the regression model. CatBoost is a gradient boosting decision tree method specifically designed for structured data with numerous categorical features. In our framework, all engineered features, including visual descriptors, user attributes, metadata, and temporal signals, are concatenated to form a comprehensive input vector \mathbf{x}_i for each post i .

During training, CatBoost utilizes category-aware encoding strategies to transform categorical features such as user identifiers, video format, music title, and posting time. The regression objective is formulated as the minimization of the Huber loss function, which can be written as:

$$\mathcal{L}_\delta(y, \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2, & \text{if } |y - \hat{y}| \leq \delta \\ \delta|y - \hat{y}| - \frac{1}{2}\delta^2, & \text{otherwise} \end{cases} \quad (3)$$

where y denotes the ground-truth popularity score and \hat{y} is the predicted value. The parameter δ determines the transition point between quadratic and linear behavior, thereby improving robustness to outliers while maintaining sensitivity to regular samples.

We apply five-fold cross-validation by splitting the training set into five subsets and iteratively using four for training and one for validation. For each fold k , the CatBoost regressor $f^{(k)}$ is trained to learn the mapping:

$$\hat{y}_i^{(k)} = f^{(k)}(\mathbf{x}_i) \quad (4)$$

The final prediction \hat{y}_i is calculated by averaging the outputs of all folds:

$$\hat{y}_i = \frac{1}{K} \sum_{k=1}^K \hat{y}_i^{(k)} \quad (5)$$

where $K = 5$ in our experiments.

4 EXPERIMENT

4.1 Dataset

The SMPD-Video dataset comprises 6,000 video posts from 4,500 users, spanning 24 months and covering 120 categories. Each post is annotated with detailed metadata, including user statistics, video attributes, posting time, location, captions, and over 40,000 unique tags. The dataset supports multimodal video popularity prediction with comprehensive visual, textual, and user-related information.

To address the substantial variance in engagement metrics across different videos, where view counts can range from zero to millions, the dataset employs a logarithmic transformation to standardize the popularity labels. Following established practices [35] in social media analysis, the normalized popularity score is computed as:

$$s = \log_2 \frac{r}{d} + 1 \quad (6)$$

where s represents the transformed popularity score, r denotes the raw view count, and d indicates the number of days elapsed since publication. This normalization approach effectively reduces the impact of extreme values while preserving the relative ranking of content popularity, thereby facilitating more stable model training and evaluation.

Figure 2 presents the distribution of user post counts in the SMPD-Video dataset. Most users contribute only a small number of posts, while a minority generate many posts. This long-tailed pattern reflects typical user activity on social media platforms. In comparison with SMPD-Image [33], the SMPD-Video dataset covers fewer posts per user, as the collection process yields a more limited number of videos for each user. As a result, the prediction task in the video track places greater emphasis on the features of individual videos and their content, rather than relying on extensive user history.

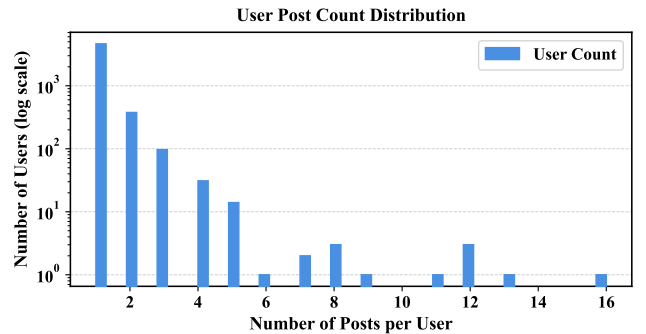


Figure 2: Distribution of user post counts in the SMPD-Video dataset.

4.2 Evaluation Metrics

The effectiveness of our approach is evaluated using the Mean Absolute Percentage Error (MAPE). As a scale-invariant metric, MAPE is suitable for measuring prediction accuracy across diverse popularity values. For a set of n samples, let y_i denote the ground-truth popularity and \hat{y}_i the predicted value for the i -th video. The MAPE is defined as

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (7)$$

MAPE provides a scale-invariant measure of accuracy, making it particularly suitable for popularity prediction tasks with large value ranges.

4.3 Main Results

4.3.1 Overall Performance. To assess the effectiveness of the proposed approach, we report the final performance on the official evaluation set using the Mean Absolute Percentage Error (MAPE) as the primary metric. Our solution achieves a MAPE of **0.1754**. This result demonstrates the superiority of the designed model architecture and feature engineering pipeline in capturing the intricate relationships underlying social media popularity prediction.

4.3.2 Ablation Study. We conduct a comprehensive ablation study to clarify the contribution of each model component and design choice. **Table 1** summarizes the results as we systematically remove or alter each major module and feature group. The full model achieves a MAPE of 0.1754, establishing a strong performance baseline. Removing video embedding features increases the MAPE to 0.1818, which demonstrates the crucial role of visual information for popularity prediction. Without text embedding features, the MAPE rises to 0.1810, confirming that semantic information from textual content supports accurate predictions. When we exclude tag popularity statistics, the MAPE reaches 0.1785, indicating that social trend signals provide valuable context for the model. Removing video metadata features results in a MAPE of 0.1782, while eliminating temporal features leads to a MAPE of 0.1758. These results show that both technical attributes and temporal signals offer additional, though moderate, improvements in prediction accuracy.

We also investigate the impact of preprocessing and model aggregation strategies. When we remove the outlier removal procedure, the MAPE increases to 0.1839, which highlights the necessity of robust data cleaning to handle noisy social media inputs. Eliminating the K-fold ensemble strategy results in a MAPE of 0.1786, underscoring the performance gains achieved through model aggregation. Among all feature groups, user-related features prove most influential. Excluding these features causes the MAPE to surge to 0.3010, revealing that user engagement history and social network characteristics represent primary predictors for content virality. This dramatic performance drop illustrates how information about user behavior and connections enables the model to capture underlying popularity dynamics that other features cannot provide.

Overall, the ablation results demonstrate that the model’s success arises from the interplay of multimodal feature extraction, effective data preprocessing, and ensemble learning. Each component contributes to performance improvements, while user-centric and

semantic features deliver especially significant benefits. These findings highlight the importance of designing a holistic approach that integrates diverse sources of information and leverages advanced data processing techniques to maximize predictive accuracy.

Table 1: Ablation study results on the validation set.

Methods	MAPE
MVP	0.1754
w/o Video Embedding	0.1818
w/o Text Embedding	0.1810
w/o Filter	0.1839
w/o User Profile	0.3010
w/o Video Metadata	0.1782
w/o Tag Popularity	0.1785
w/o Time	0.1758
w/o K-Fold	0.1786

4.3.3 Feature Importance. To understand which features contribute most to the model’s prediction, we analyze the feature importance scores output by CatBoost. **Table 2** lists the top 20 features ranked by their importance.

Among these features, user engagement indicators such as *LikeCount*, *VideoCount*, and *HeartCount* are assigned the highest importance, suggesting that historical activity and user popularity play a substantial role in video popularity prediction. Additionally, several user-related and content-based features, including *FollowerCount* and *SuggestedWordsLen*, demonstrate considerable influence on the model’s output. It is noteworthy that latent features extracted by dimensionality reduction methods, such as the *svd_mode* series and *video_embedding* vectors, also appear frequently in the top ranks, indicating that the model benefits from both explicit and implicit representations of multimodal data.

The distribution of importance scores reflects the complementary effect of combining user metadata, textual information, and deep video embeddings within a unified framework. These findings confirm that a balanced integration of heterogeneous features can enhance predictive performance for the popularity estimation task. Furthermore, the relative importance of user-related features highlights the continued influence of user history and social reach in content dissemination on social platforms.

Table 2: Importance of different features.

Rank	Feature	Importance	Rank	Feature	Importance
1	LikeCount	13.775	11	FollowingCount	1.739
2	VideoCount	12.392	12	svd_mode_t_11	1.137
3	HeartCount	12.033	13	svd_mode_t_16	1.130
4	FollowerCount	9.986	14	svd_mode_t_1	1.113
5	SuggestedWordsLen	4.184	15	svd_mode_4	1.009
6	AvgTagPopularity	3.822	16	svd_mode_7	0.945
7	svd_mode_0	2.621	17	PostContentLen	0.942
8	PostCount	2.529	18	video_embedding_15	0.887
9	PostLocation	2.492	19	svd_mode_t_18	0.868
10	svd_mode_t_2	2.058	20	video_embedding_19	0.820

4.4 Analysis

To further explore the alignment between our model's predictions and the true popularity scores, we conduct a statistical analysis on the validation set. Figure 3 illustrates the histograms and density distributions for both predicted and actual labels.

The predicted scores demonstrate a strong correspondence to the empirical distribution of the ground-truth labels, effectively capturing the unimodal and skewed nature of the data. Notably, the peak density of the predictions occurs in a similar range as the true scores (approximately 5 to 10), indicating that the dominant pattern present in the data is accurately learned by the model. A slight smoothing in the prediction curve is observed, which is likely a result of ensemble inference and regularization strategies designed to mitigate overfitting.

Despite this overall alignment, the model shows relatively conservative estimates at the distribution extremes. While predictions span the entire label range, there is a tendency to underestimate scores in the low-popularity (<2.5) and high-popularity (>15) intervals. This phenomenon is common in regression tasks involving long-tailed distributions and can be partially attributed to the imbalance of training samples in these regions, as reflected in the ground-truth histogram. To address this limitation, future work may consider techniques such as label distribution smoothing or cost-sensitive re-weighting to improve model performance on rare cases.

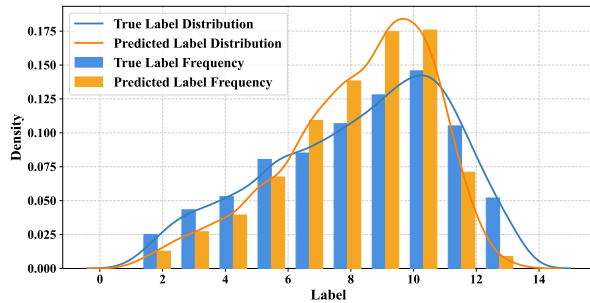


Figure 3: Histogram and kernel density estimation of predicted vs. ground-truth label distributions on the validation set.

Overall, the predicted results exhibit a high degree of statistical fidelity to the actual label distribution. The model is capable of producing scores with meaningful variance and reliable calibration, supporting its potential for large-scale popularity prediction in realistic social media scenarios.

5 CONCLUSION AND FUTURE WORK

This work presents a unified and robust framework for predicting the popularity of social media videos, leveraging comprehensive multimodal feature extraction, advanced machine learning models, and ensemble strategies. Through careful integration of visual, textual, user, and structural information, the proposed approach effectively captures the complex patterns underlying social engagement dynamics. Experimental results demonstrate that incorporating

domain-specific features and explicitly modeling user behavior can substantially improve predictive accuracy.

The MVP framework exhibits strong potential for cross-platform transferability, as the core multimodal feature extraction and ensemble learning components can be adapted to different social media environments with minimal architectural modifications. The modular design facilitates deployment across various content types and user demographics, making it suitable for real-world applications in content recommendation systems and marketing analytics.

Nonetheless, challenges such as semantic inconsistencies across modalities and the dynamic nature of social media content remain open for further exploration [3]. Future research may focus on developing more adaptive multimodal fusion techniques, as well as end-to-end architectures capable of leveraging temporal patterns and contextual cues. Moreover, enhancing the interpretability and robustness of prediction systems through improved alignment and representation learning is a promising direction.

In future work, the MVP framework can be extended to support online popularity forecasting under streaming settings, or adapted to incorporate graph-based user interaction structures and multi-task objectives, such as virality classification and trend forecasting. Graph neural networks could capture complex social network dynamics and user influence propagation patterns, while multi-task learning approaches might simultaneously predict engagement metrics, content lifespan, and audience demographics. Additionally, developing real-time deployment strategies for continuous model updating and adaptive threshold adjustment would enhance practical applicability in dynamic social media environments.

In summary, this study provides a meaningful foundation for further multimodal social media analysis and offers valuable insights into building more accurate and generalizable content popularity prediction models.

References

- [1] Hervé Abdi and Lynne J Williams. 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics* 2, 4 (2010), 433–459.
- [2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. ViViT: A Video Vision Transformer. *arXiv:2103.15691* [cs] doi:10.48550/arXiv.2103.15691
- [3] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41, 2 (2018), 423–443.
- [4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is Space-Time Attention All You Need for Video Understanding? *arXiv:2102.05095* [cs] doi:10.48550/arXiv.2102.05095
- [5] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.
- [6] Zhangtao Cheng, Jienan Zhang, Xovee Xu, Goce Trajcevski, Ting Zhong, and Fan Zhou. 2024. Retrieval-augmented hypergraph for multimodal social media popularity prediction. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*. 445–455.
- [7] Cisco. 2020. Cisco Annual Internet Report - Cisco Annual Internet Report (2018–2023) White Paper. <https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html>.
- [8] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. 2018. CatBoost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363* (2018).
- [9] Wenhao Hu, Weilong Chen, Weimin Yuan, Yan Wang, Shimin Cai, and Yanru Zhang. 2024. Dual-Stream Pre-Training Transformer to Enhance Multimodal Learning for Social Media Prediction. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 11450–11456.
- [10] Yangliu Hu, Zikai Song, Na Feng, Yawei Luo, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. 2025. Sf2t: Self-supervised fragment finetuning of video-llms for fine-grained understanding. In *Proceedings of the Computer Vision and Pattern*

- Recognition Conference. 29108–29117.
- [11] Xin Lai, Yihong Zhang, and Wei Zhang. 2020. HyFea: Winning Solution to Social Media Popularity Prediction for Multimedia Grand Challenge 2020. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*. Association for Computing Machinery, New York, NY, USA, 4565–4569. doi:10.1145/3394171.3416273
 - [12] Wenbing Li, Hang Zhou, Junqing Yu, Zikai Song, and Wei Yang. 2024. Coupled mamba: Enhanced multi-modal fusion with coupled state space model. *arXiv preprint arXiv:2405.18014* (2024).
 - [13] Haixu Liu, Wenning Wang, Haoxiang Zheng, Penghao Jiang, Qirui Wang, Ruiqing Yan, and Qiuzhuang Sun. 2025. Multi-Modal Video Feature Extraction for Popularity Prediction. arXiv:2501.01422 [cs] doi:10.48550/arXiv.2501.01422
 - [14] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. 2019. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487* (2019).
 - [15] Run Luo, Zikai Song, Lintao Ma, Jinlin Wei, Wei Yang, and Min Yang. 2024. Diffusiontrack: Diffusion model for multi-object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 3991–3999.
 - [16] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. 2022. X-CLIP: End-to-End Multi-grained Contrastive Learning for Video-Text Retrieval. arXiv:2207.07285 [cs] doi:10.48550/arXiv.2207.07285
 - [17] Shijian Mao, Wudong Xi, Lei Yu, Gaotian Lü, Xingxing Xing, Xingchen Zhou, and Wei Wan. 2023. Enhanced CatBoost with Stacking Features for Social Media Prediction. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*. Association for Computing Machinery, New York, NY, USA, 9430–9435. doi:10.1145/3581783.3612839
 - [18] Mayank Meghawat, Satyendra Yadav, Debanjan Mahata, Yifang Yin, Rajiv Ratn Shah, and Roger Zimmermann. 2018. A Multimodal Approach to Predict Social Media Popularity. arXiv:1807.05959 [cs] doi:10.48550/arXiv.1807.05959
 - [19] Antoine Miech, Ivan Laptev, and Josef Sivic. 2018. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516* (2018).
 - [20] Thong Nguyen, Yi Bin, Junbin Xiao, Leigang Qu, Yicong Li, Jay Zhangjie Wu, Cong-Duy Nguyen, See-Kiong Ng, and Luu Anh Tuan. 2024. Video-Language Understanding: A Survey from Model Architecture, Model Training, and Data Perspectives. arXiv:2406.05615 [cs] doi:10.48550/arXiv.2406.05615
 - [21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmlR, 8748–8763.
 - [22] Zikai Song, Run Luo, Lintao Ma, Ying Tang, Yi-Ping Phoebe Chen, Junqing Yu, and Wei Yang. 2025. Temporal Coherent Object Flow for Multi-Object Tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 6978–6986.
 - [23] Zikai Song, Run Luo, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. 2023. Compact transformer tracker with correlative masked modeling. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 2321–2329.
 - [24] Zikai Song, Ying Tang, Run Luo, Lintao Ma, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. 2024. Autogenic language embedding for coherent point tracking. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 2021–2030.
 - [25] Zikai Song, Zhiwen Wan, Wei Yuan, Ying Tang, Junqing Yu, and Yi-Ping Phoebe Chen. 2021. Distractor-aware tracker with a domain-special optimized benchmark for soccer player tracking. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*. 276–284.
 - [26] Zikai Song, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. 2022. Transformer tracking with cyclic shifting window attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8791–8800.
 - [27] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*. 7464–7473.
 - [28] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. VideoMAE: Masked Autoencoders Are Data-Efficient Learners for Self-Supervised Video Pre-Training. https://arxiv.org/abs/2203.12602v3.
 - [29] Mingsheng Tu, Tianjiao Wan*, Qisheng Xu, Xinhao Jiang, Kele Xu, and Cheng Yang. 2024. Higher-Order Vision-Language Alignment for Social Media Prediction. In *Proceedings of the 32nd ACM International Conference on Multimedia*. ACM, Melbourne VIC Australia, 11457–11463. doi:10.1145/3664647.3688999
 - [30] Caroline Violot, Tuğrulcan Elmas, Igor Bilogrevic, and Mathias Humbert. 2024. Shorts vs. regular videos on YouTube: a comparative analysis of user engagement and content creation trends. In *Proceedings of the 16th ACM Web Science Conference*. 213–223.
 - [31] Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, et al. 2022. Language models with image descriptors are strong few-shot video-language learners. *Advances in Neural Information Processing Systems* 35 (2022), 8483–8497.
 - [32] Bo Wu, Wen-Huang Cheng, Peiye Liu, Bei Liu, Zhaoyang Zeng, and Jiebo Luo. 2019. Smp challenge: An overview of social media prediction challenge 2019. In *Proceedings of the 27th ACM International Conference on Multimedia*. 2667–2671.
 - [33] Bo Wu, Peiye Liu, Wen-Huang Cheng, Bei Liu, Zhaoyang Zeng, Jia Wang, Qiushi Huang, and Jiebo Luo. 2023. SMP Challenge: An overview and analysis of social media prediction challenge. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9651–9655.
 - [34] Bo Wu, Peiye Liu, Qiushi Huang, Zhaoyang Zeng, Jia Wang, Bei Liu, Jiebo Luo, and Wen-Huang Cheng. 2024. SMP Challenge Summary: Social Media Prediction Challenge. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 11442–11444.
 - [35] Bo Wu, Tao Mei, Wen-Huang Cheng, and Yongdong Zhang. 2016. Unfolding temporal dynamics: Predicting social media popularity using multi-scale temporal decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.
 - [36] Kele Xu, Zhimin Lin, Jianqiao Zhao, Peicang Shi, Wei Deng, and Huaimin Wang. 2020. Multimodal deep learning for social media popularity prediction with attention mechanism. In *Proceedings of the 28th ACM International Conference on Multimedia*. 4580–4584.
 - [37] Hang Zhou, Jiale Cai, Yuteng Ye, Yonghui Feng, Chenxing Gao, Junqing Yu, Zikai Song, and Wei Yang. 2025. Video anomaly detection with motion and appearance guided patch diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 10761–10769.
 - [38] Yaochen Zhu, Jiayi Xie, and Zhenzhong Chen. 2020. Predicting the Popularity of Micro-videos with Multimodal Variational Encoder-Decoder Framework. https://arxiv.org/abs/2003.12724v1. doi:10.1109/TMM.2021.3120537

References

- [1] Hervé Abdi and Lynne J Williams. 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics* 2, 4 (2010), 433–459.
- [2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. ViViT: A Video Vision Transformer. arXiv:2103.15691 [cs] doi:10.48550/arXiv.2103.15691
- [3] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41, 2 (2018), 423–443.
- [4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is Space-Time Attention All You Need for Video Understanding? arXiv:2102.05095 [cs] doi:10.48550/arXiv.2102.05095
- [5] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.
- [6] Zhanqiao Cheng, Jienan Zhang, Xovee Xu, Goce Trajcevski, Ting Zhong, and Fan Zhou. 2024. Retrieval-augmented hypergraph for multimodal social media popularity prediction. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*. 445–455.
- [7] Cisco. 2020. Cisco Annual Internet Report - Cisco Annual Internet Report (2018–2023) White Paper. https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html.
- [8] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. 2018. CatBoost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363* (2018).
- [9] Wenhao Hu, Weilong Chen, Weimin Yuan, Yan Wang, Shimin Cai, and Yanru Zhang. 2024. Dual-Stream Pre-Training Transformer to Enhance Multimodal Learning for Social Media Prediction. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 11450–11456.
- [10] Yangliu Hu, Zikai Song, Na Feng, Yawei Luo, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. 2025. Sf2t: Self-supervised fragment finetuning of video-llms for fine-grained understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 29108–29117.
- [11] Xin Lai, Yihong Zhang, and Wei Zhang. 2020. HyFea: Winning Solution to Social Media Popularity Prediction for Multimedia Grand Challenge 2020. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*. Association for Computing Machinery, New York, NY, USA, 4565–4569. doi:10.1145/3394171.3416273
- [12] Wenbing Li, Hang Zhou, Junqing Yu, Zikai Song, and Wei Yang. 2024. Coupled mamba: Enhanced multi-modal fusion with coupled state space model. *arXiv preprint arXiv:2405.18014* (2024).
- [13] Haixu Liu, Wenning Wang, Haoxiang Zheng, Penghao Jiang, Qirui Wang, Ruiqing Yan, and Qiuzhuang Sun. 2025. Multi-Modal Video Feature Extraction for Popularity Prediction. arXiv:2501.01422 [cs] doi:10.48550/arXiv.2501.01422
- [14] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. 2019. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487* (2019).
- [15] Run Luo, Zikai Song, Lintao Ma, Jinlin Wei, Wei Yang, and Min Yang. 2024. Diffusiontrack: Diffusion model for multi-object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 3991–3999.
- [16] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. 2022. X-CLIP: End-to-End Multi-grained Contrastive Learning for Video-Text Retrieval.

- arXiv:2207.07285 [cs] doi:10.48550/arXiv.2207.07285
- [17] Shijian Mao, Wudong Xi, Lei Yu, Gaotian Lü, Xingxing Xing, Xingchen Zhou, and Wei Wan. 2023. Enhanced CatBoost with Stacking Features for Social Media Prediction. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*. Association for Computing Machinery, New York, NY, USA, 9430–9435. doi:10.1145/3581783.3612839
 - [18] Mayank Meghawat, Satyendra Yadav, Debanjan Mahata, Yifang Yin, Rajiv Ratn Shah, and Roger Zimmermann. 2018. A Multimodal Approach to Predict Social Media Popularity. arXiv:1807.05959 [cs] doi:10.48550/arXiv.1807.05959
 - [19] Antoine Miech, Ivan Laptev, and Josef Sivic. 2018. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516* (2018).
 - [20] Thong Nguyen, Yi Bin, Junbin Xiao, Leigang Qu, Yicong Li, Jay Zhangjie Wu, Cong-Duy Nguyen, See-Kiong Ng, and Luu Anh Tuan. 2024. Video-Language Understanding: A Survey from Model Architecture, Model Training, and Data Perspectives. arXiv:2406.05615 [cs] doi:10.48550/arXiv.2406.05615
 - [21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmlR, 8748–8763.
 - [22] Zikai Song, Run Luo, Lintao Ma, Ying Tang, Yi-Ping Phoebe Chen, Junqing Yu, and Wei Yang. 2025. Temporal Coherent Object Flow for Multi-Object Tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 6978–6986.
 - [23] Zikai Song, Run Luo, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. 2023. Compact transformer tracker with correlative masked modeling. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 2321–2329.
 - [24] Zikai Song, Ying Tang, Run Luo, Lintao Ma, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. 2024. Autogenic language embedding for coherent point tracking. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 2021–2030.
 - [25] Zikai Song, Zhiwen Wan, Wei Yuan, Ying Tang, Junqing Yu, and Yi-Ping Phoebe Chen. 2021. Distractor-aware tracker with a domain-special optimized benchmark for soccer player tracking. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*. 276–284.
 - [26] Zikai Song, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. 2022. Transformer tracking with cyclic shifting window attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8791–8800.
 - [27] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*. 7464–7473.
 - [28] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. VideoMAE: Masked Autoencoders Are Data-Efficient Learners for Self-Supervised Video Pre-Training. <https://arxiv.org/abs/2203.12602v3>.
 - [29] Mingsheng Tu, Tianjiao Wan*, Qisheng Xu, Xinhao Jiang, Kele Xu, and Cheng Yang. 2024. Higher-Order Vision-Language Alignment for Social Media Prediction. In *Proceedings of the 32nd ACM International Conference on Multimedia*. ACM, Melbourne VIC Australia, 11457–11463. doi:10.1145/3664647.3688999
 - [30] Caroline Violot, Tuğrulcan Elmas, Igor Bilogrevic, and Mathias Humbert. 2024. Shorts vs. regular videos on YouTube: a comparative analysis of user engagement and content creation trends. In *Proceedings of the 16th ACM Web Science Conference*. 213–223.
 - [31] Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, et al. 2022. Language models with image descriptors are strong few-shot video-language learners. *Advances in Neural Information Processing Systems* 35 (2022), 8483–8497.
 - [32] Bo Wu, Wen-Huang Cheng, Peiye Liu, Bei Liu, Zhaoyang Zeng, and Jiebo Luo. 2019. Smp challenge: An overview of social media prediction challenge 2019. In *Proceedings of the 27th ACM International Conference on Multimedia*. 2667–2671.
 - [33] Bo Wu, Peiye Liu, Wen-Huang Cheng, Bei Liu, Zhaoyang Zeng, Jia Wang, Qiushi Huang, and Jiebo Luo. 2023. SMP Challenge: An overview and analysis of social media prediction challenge. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9651–9655.
 - [34] Bo Wu, Peiye Liu, Qiushi Huang, Zhaoyang Zeng, Jia Wang, Bei Liu, Jiebo Luo, and Wen-Huang Cheng. 2024. SMP Challenge Summary: Social Media Prediction Challenge. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 11442–11444.
 - [35] Bo Wu, Tao Mei, Wen-Huang Cheng, and Yongdong Zhang. 2016. Unfolding temporal dynamics: Predicting social media popularity using multi-scale temporal decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.
 - [36] Kele Xu, Zhimin Lin, Jianqiao Zhao, Peicang Shi, Wei Deng, and Huaimin Wang. 2020. Multimodal deep learning for social media popularity prediction with attention mechanism. In *Proceedings of the 28th ACM International Conference on Multimedia*. 4580–4584.
 - [37] Hang Zhou, Jiale Cai, Yuteng Ye, Yonghui Feng, Chenxing Gao, Junqing Yu, Zikai Song, and Wei Yang. 2025. Video anomaly detection with motion and appearance guided patch diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 10761–10769.
 - [38] Yaochen Zhu, Jiayi Xie, and Zhenzhong Chen. 2020. Predicting the Popularity of Micro-videos with Multimodal Variational Encoder-Decoder Framework. <https://arxiv.org/abs/2003.12724v1>. doi:10.1109/TMM.2021.3120537